

GBREL.TXT Genetic Sequence Data Bank 15 February 1997 NCBI-GenBank Flat File Release 99.0 Distribution CD-ROM Release Notes 1192505 loci, 786898138 bases, from 1192505 reported sequences This document describes the data written on GenBank flat file distribution CD-ROMs. If you have any questions or comments about the data bank, the CD-ROM, or this document, please contact NCBI via email at info@ncbi.nlm.nih.gov or: GenBank National Center for Biotechnology Information National Library of Medicine, 38A, 8N805 8600 Rockville Pike Bethesda, MD 20894 USA Phone: (301) 496-2475 Fax: (301) 480-9241 1. INTRODUCTION 1.1 Release 99.0 The National Center for Biotechnology Information (NCBI) at the National Library of Medicine (NLM), National Institutes of Health (NIH) is responsible for producing and distributing the GenBank Sequence Data Bank. NCBI handles all GenBank direct submission data and authors are advised to use the address below. Submitters are encouraged to use the free Sequin software package for sending sequence data, or the newly developed World Wide Web submission form. See Section 1.5 below for details.

***** The address for direct submissions to GenBank is: GenBank Submissions National Center for Biotechnology Information Bldg 38A, Rm. 8N-803 8600 Rockville Pike Bethesda, MD 20894 E-MAIL: gb-sub@ncbi.nlm.nih.gov Updates and changes to existing GenBank records: E-MAIL: update@ncbi.nlm.nih.gov URL for the new GenBank submission tool - BankIt - on the World Wide Web: <http://www.ncbi.nlm.nih.gov/> (see Section 1.5 for additional details about submitting data to GenBank.)

***** Release 99.0 is a release of sequence data by NCBI in the GenBank flat file format. It contains a large number of entries produced at the NLM, derived from scanning the biomedical literature. GenBank contains sequences submitted directly to the database by authors as well as entries created at NLM derived from scanning the biomedical literature. Over 325,000 articles per year from 3400 journals are scanned for sequence data. They are supplemented by journals in plant and veterinary sciences through a collaboration with the National Agricultural Library. GenBank is a component of a tripartite, international collaboration of sequence databases in the U.S., Europe, and Japan. The collaborating databases in Europe are the European Molecular Biology Laboratory (EMBL) at Hinxton Hall, UK, and the DNA Database of Japan (DDBJ) in Mishima, Japan. Sequence data is also incorporated from the Genome Sequence Data Base (GSDB), Santa Fe, NM. Patent sequences are incorporated through arrangements with the U.S. Patent and Trademark Office, and via the collaborating international databases from other international patent offices. The database is converted to various output formats, including the Flat File and Abstract Syntax Notation 1 (ASN.1) versions. The ASN.1 form of the data is included on the Entrez: Sequences CD-ROM and is also available, as is the flat file, by anonymous FTP to 'ncbi.nlm.nih.gov'. 1.2 Cutoff Date This full release, 99.0, incorporates data available to the databases as of February 7, 1997. For more recent data, users are advised to download the update files by anonymous FTP to 'ncbi.nlm.nih.gov' or to search the updates via the e-mail server. For instructions on the use of the e-mail server, send mail message with the word 'help' in it to: retrieve@ncbi.nlm.nih.gov 1.3 Important Changes in Release 99.0 1.3.1 New EST files Due to the growth in the number of EST sequences, the EST division is now being split into 11 pieces. 1.4 Upcoming Changes No changes to the GenBank Flatfile release expected for Release 100.0 . 1.5 Request for Direct Submission of Sequence Data A successful GenBank requires that the data enter the database as soon as possible after publication, that the annotations be as complete as possible, and that the sequence and annotation data be accurate. All three of these requirements are best met if authors of sequence data submit their data directly to GenBank in a usable form. It is especially important that these submissions be in computer-readable form. GenBank must rely on

direct author submission of data to ensure that it achieves its goals of completeness, accuracy, and timeliness. To assist researchers in entering their own sequence data, GenBank provides a WWW submission tool called BankIt, as well as a stand-alone software package called Sequin. BankIt and Sequin are both easy-to-use programs that enable authors to enter a sequence, annotate it, and submit it to GenBank. Through the international collaboration of DNA sequence databases, GenBank submissions are forwarded daily for inclusion in the EMBL and DDBJ databases. SEQUIN. Sequin is an interactive, graphically-oriented program based on screen forms and controlled vocabularies that guides you through the process of entering your sequence and providing biological and bibliographic annotation. Intended as an alternative to the older Authorin program, Sequin is designed to simplify the sequence submission process, and to provide increased data handling capabilities to accommodate very long sequences, complex annotations, and robust error checking. E-mail the completed submission file to : gb-sub@ncbi.nlm.nih.gov Sequin is currently provided as a beta-test version, and runs on Macintosh, PC/Windows, UNIX and VMS computers. It is available by anonymous ftp from ncbi.nlm.nih.gov , login as anonymous and use your e-mail address as the password. It is located in the sequin directory. BANKIT. BankIt provides a simple forms approach for submitting your sequence and descriptive information to GenBank. Your submission will be submitted directly to GenBank via the World Wide Web, and immediately forwarded for inclusion in the EMBL and DDBJ databases. BankIt may be used with Netscape clients for Unix, Macs, and PCs, the Mosaic client for Unix, and the MacWeb client for Macs. You can access BankIt from GenBank's home page: <http://www.ncbi.nlm.nih.gov/> AUTHORIN. Authorin is no longer the primary means for submitting sequences to GenBank, and is no longer being distributed by NCBI. For submitters who already have this program, however, we do continue to accept Authorin submissions. For those who are unable to use Sequin or BankIt, GenBank has an ASCII text electronic data submission form. This form is standardized among EMBL, DDBJ, GenBank, PIR, MIPS, and JIPID. The GenBank Data Submission Form (located in the file GBDAT.FRM) can be used to submit your sequence and annotations. Electronic mail submissions should go to: gb-sub@ncbi.nlm.nih.gov. Direct mail on floppy disk should go to: GenBank Submissions National Center for Biotechnology Information Bldg. 38A, Rm 8N-803 8600 Rockville Pike Bethesda, MD 20894 If you have questions about GenBank submissions or any of the data submission tools, contact NCBI at: info@ncbi.nlm.nih.gov or 301-496-2475.

1.6 Organization of This Document The second section describes the contents of the CD-ROM files. The third section illustrates the formats of the CD-ROM files. The fourth section describes other versions of the data, the fifth section identifies known problems, and the sixth contains administrative details and ordering information.

2. ORGANIZATION OF CD-ROM FILES

2.1 CD-ROM Format